An Exploration of Agrotech

February 24, 2022

Registration number:	2100374
Project:	Agrotech
Link to GitHub:	https://github.com/rigovm101/CE888_DataScience/tree/main/Project

Executive summary (max. 250 words)	78
Introduction (max. 600 words)	164
Data (max. 500 words/dataset)	361
Methodology (max. 600 words)	124
Conclusions (max. 500 words)	74
Total word count	821

Contents

Abstract

The goal of this project is to create a model able to predict the size of crops using weather data. This will allow farmers to maximize profits and supermarkets to cut down on food waste. We'll be using the Agrotech dataset to train and test this model. After cleaning and pre-processing the dataset weather data was incorporated into the given features. The data is ready to be implemented into a multi-label regressor which will be evaluated using RMSE.

1 Introduction

Optimization is a key part of business, everyone wants to maximize profits while minimizing costs. There are some sectors in which controlling the environment is simple, but in the case of the farming industry, almost everyone depends on appropiate weather conditions. Not only that, but supply and demand also adds an uncertainty factor when trying to optimize this industry.

The main motivation behind this project is to use Machine Learning to give farmers a tool to predict the size of their crops in order to communicte better with supermarkets on their stock. This will allow supermarkets to better reduce the food waste in their stores. As noted by a study done in Brazil, one of the main causes for food waste is ”inneffective stock control management”. [2] A report here in the UK also estimates that around 40.7% of the fresh products gets wasted or not completely used [1]. We believe Machine Learning can give farmers useful information to both plan their produce and maximize profits.

2 Data

For this project we will be working with the Agrotech dataset, which was kindly provided by the project supervisor. Given the sensitivity of this data, it will be kept separate from the main repository. This dataset contains information about crops and weather. Table 1 contains more information about what each specific sheet contains.

Plants	Information about the plants
Flight Dates	Information about the flights
Planting	Additional information about the plants
Weather	Historical weather information

Table 1: Table with the information per sheet Agrotech

The sheet *Plants* was the one which contained the actual labels to use, which are *Head Weight, Polar Diameter and radial diameter.* These columns are of type Float and have some NULL values. Being the labe values, we removed these rows. The last column called *Remove* consisted of marks for rows to be deleted. We removed those rows and then dropped the column.

The *Flight Dates* sheet was very simple to process, we just filled in the missing dates in the *Plants* sheet with the dates given by this sheet. The *Planting* sheet contained some additional information about the plants. But we were told by the project supervisor that this sheet contained a mix of two different spreadsheets, so we dropped some rows which contained data in the wrong format. Once the sheet was cleaned, we filled in missing *Plant Dates* and added some additional features such as *Volume Planted.*

Finally we were left with the *Weather* sheet. For us to be able to use information from this sheet in the main DataFrame we were building, we implemented two functions. The first one returned a Series with an average of the weather conditions a plant went through between two given dates. The second function also returned a Series with weather information, but with data from the previous year. For the initial model testing we will encode in the DataFrame information from the weather between *Plant Date* and *Flight Date* to predict the final size of the crop at *Check Date.* Depending on the result we may change this weather information to get the weather conditions between *Plant Date* and *Check Date* from the previous year, also with the possibility to retrieve information from more than one year in the past.

After the processing of the data, we analyzed the distribution of the data through a Histogram, illustrated in Figure 1.

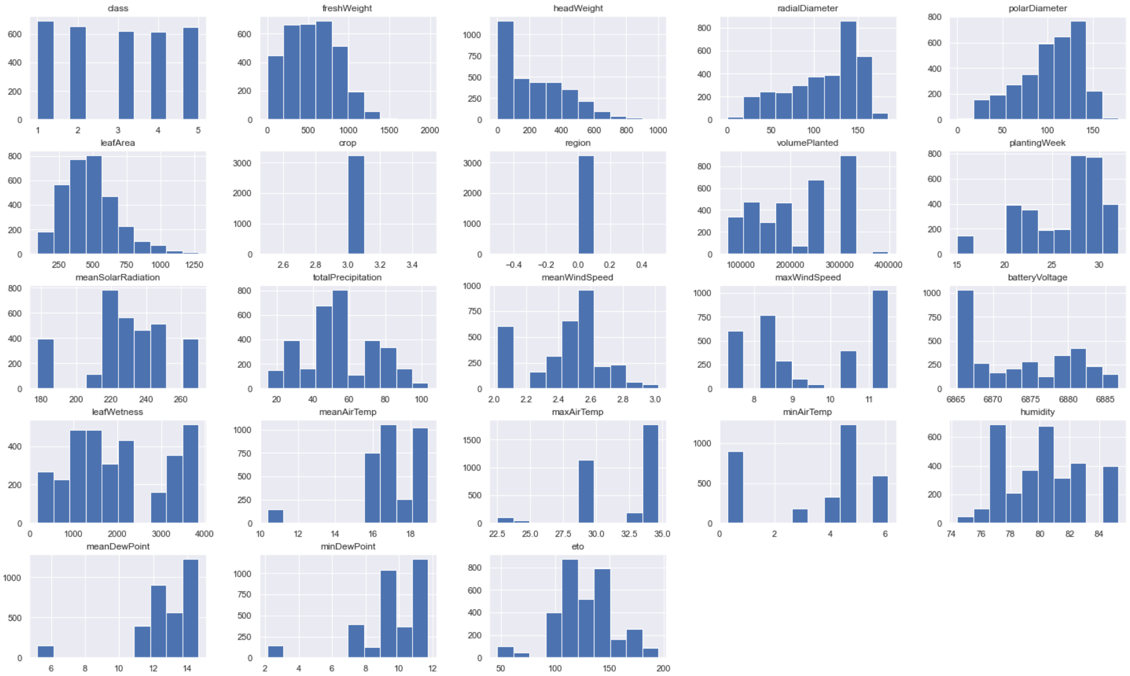

Figure 1: Histograms of all the features

3 Methodology

After analyzing the data proided, we have identified that a multi-label linear regressor will probably allow us to construct a useful model. Specifically, we plan to use the *MultiOutputRegressor* class from the *ScikitLearn* library, which allows us to use any regressor as a multi-label regressor. First we will implement Linear Regression, with the possibility to later on implement Decision Tree Regressor.

For training the model, we will use the *ScikitLearn* tools to split our data into training and validation, using a 70% split for training and 30% for validation. Being a regression problem, we will use Mean Squared Error (MSE) to measure the error during the training phase, while using the Root Mean Squared Error (RMSE) to measure the performance during the valudation phase.

4 Conclusions

While the weather is never the same, we believe it will be a useful feature to help us correctly estimate the size of the crops for better planning for both the farmers and the supermarkets. After processing the data we are confident that a useful model will be trainned and perform well. The team is open to testing other Machine Learning tools and algorithms and deleting features that turn out not to be useful.

References

[1] N. Baker, S. Popay, J. Bennett, and M. Kneafsey. Net yield efficiency: Comparing salad and vegetable waste between community supported agriculture and supermarkets in the uk. *Journal of Agriculture, Food Systems, and Community Development*, 8(4):179, Winter 2019. Copyright - Copyright New Leaf Associates, Inc. Winter/Spring 2019; Last update - 2019-06-19.

[2] M. W. Eluiza Alberto de, d. N. Caroline Rodrigues, d. F. Michele Gasparoto Moreira Teixeira, and M. V. Mayra. Food waste: an exploratory investigation of causes, practices and consequences perceived by brazilian supermarkets and restaurants. *British Food Journal*, 124(3):1022–1045, 2022. Copyright - © Emerald Publishing Limited 2021; Last update - 2022-02-07; SubjectsTermNotLitGenreText - Denmark; Brazil; Pakistan.